Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population

Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 1862–1865, March 1997
Evolution

Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population

ANDREW J. LEIGH BROWN*

Centre for HIV Research, Institute of Cell, Animal and Population Biology, University of Edinburgh, EH9 3JN Edinburgh, Scotland

Communicated by Michael Clegg, University of California, Riverside, CA, December 17, 1996 (received for review August 21, 1996)

ABSTRACT **Selection is usually considered to be the dominant force controlling viral variation; the large population sizes suggest that deterministic population genetic models are appropriate. To investigate their validity for HIV, samples of *env* gene sequences were tested for departure from neutrality because of mutation–selection balance. None of the samples departed significantly when tested as nucleotide sequences. At the amino acid level, significantly elevated diversity was detected in two samples within, but not outside, the V3 loop. The effective population number has been estimated (using a phylogenetic method) to be close to 10^3. Estimates from nucleotide diversity are about 2-fold lower. The low value of the effective population number might arise from high variability in progeny number between infected cells, from the expansion in population number from a small inoculum as the virus is transmitted between hosts, or from variable selection at linked sites. These results suggest that the population genetics of HIV are best described by stochastic models.**

The precise impact of selection in determining the level and pattern of variation in the *env* gene of HIV-1 remains unclear despite extensive studies of sequence variation (reviewed in refs. 1–3). This gene encodes known targets for cytotoxic T lymphocytes and neutralizing antibodies (4–6), and it is a major determinant of cell tropism (7, 8), but there is only indirect evidence (based on a lower frequency of synonymous nucleotide substitutions than those that result in a change of the amino acid sequence in the V3 loop) that variation in *env* is maintained by selection for antigenic diversity (9, 10). To date, there has been no formal attempt to test population data of *env* sequences for departures from the expectations of selective neutrality even though nonselective explanations for *env* gene diversity have been proposed (11). In most theoretical discussions, it has been assumed that selection and mutation are the only important forces affecting viral variability; the evolution of the virus has been treated deterministically, and the population size (N) in an infected individual has been considered to be infinite (12, 13).

Although the population number of HIV is indeed very large (14, 15), data on the genetic variability of HIV-1 within an infected patient reveal features that are not explained by a deterministic model. For example, despite the consistent replicative advantage enjoyed by genomes bearing alleles conferring resistance to zidovudine, the outcome in terms of the rate at which such alleles become fixed is unpredictable. In a minority of patients, the methionine→leucine substitution at codon 41 in the reverse transcriptase domain of the *pol* gene [which confers the highest levels of zidovudine resistance (16)] is never observed (17, 18). Such unpredictability is not expected in deterministic models; they are generally not considered to be biologically realistic because of their failure to allow for sampling effects (19). In contrast, stochastic, finite population models recognize the importance of sampling in evolution. They also allow for a class of selectively neutral mutations whose fate is determined entirely by random genetic drift. The majority of synonymous nucleotide substitutions are usually considered to belong to this class whereas it has been suggested under a deterministic model that the level of synonymous variation in HIV is determined by a balance between the forward mutation rate and the rate of removal by purifying selection (mutation–selection balance) (13).

To assess the appropriateness of deterministic models for HIV populations, I have tested whether the pattern of variation in HIV-1 *env* sequences departs significantly from neutrality. The test applied can detect departures due to mutation–selection balance and those due to selection for antigenic diversity (20). No significant departure was observed overall at the level of the nucleotide sequences; this allowed estimation of the effective population number of HIV-1 within a patient, which was found to be close to 10^3, about five orders of magnitude lower than the census size for this patient.

MATERIALS AND METHODS

The data analyzed were obtained from five serial, postseroconversion plasma samples from an HIV-1-infected hemophiliac at ≈1-year intervals (years 3–7, inclusive) that were subjected to limiting dilution-nested PCR after reverse transcription of viral *env* sequences and were manually sequenced (9). A phylogenetic analysis of the sequence data has been published (21). A total of 77 sequences of the V3 region, 231 bp long, were obtained; the overall level of variation was similar to that observed in other studies (22, 23). Only plasma viral sequences were included to ensure that the sample represented the recently replicated viral population; it has been shown that the proviral population in peripheral blood is heterogeneous, containing both recently replicated and older variants (9). During this period, the patient had a declining $CD4^+$ cell count but was clinically asymptomatic. Clinical details have been published elsewhere (9). Estimates of plasma viral load, obtained in the course of the sequencing studies, ranged from 2.8×10^4 copies/ml (year 3) to 5.8×10^3 copies/ml in year 7 (L.Q. Zhang, unpublished data), suggesting an overall viral population of the order of 10^8 during the study period. A second data set of 81 amino acid sequences of the V1/V2 region of *env* (from four individual splenic white pulps from a single patient) published by Cheynier *et al.* was also analyzed (11).

Estimates of θ ($2N_e\mu$) were obtained using the program FLUCTUATE, a development of the COALESCE program described by Kuhner *et al.* (24). This was obtained from the the LAMARC package distributed by M. Kuhner and J. Felsenstein (Department of Genetics, University of Washington, Seattle). It was compiled on a Sun Sparcstation Model 10 (Sun Micro-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

0027-8424/97/941862-4$2.00/0
PNAS is available online at **http://www.pnas.org**.

Abbreviation: N_e, effective population number.

*e-mail: A.Leigh-Brown@ed.ac.uk.

Table 1. Parameters of genetic variability and a test for neutrality in sequences of the V3 region (231 bp) in five samples of plasma viral RNA from an infected hemophiliac

Sample, year	n	S_{nuc}	K_{nuc}	D_{nuc}	S_{aa}	K_{aa}	D_{aa}	S_{syn}	K_{syn}	D_{syn}
3	15	25	7.37	−0.172	19	6.47	0.438	3	1.3	1.22
4	11	21	7.89	0.457	13	6.02	1.560	4	0.87	−1.329
5	23	25	6.27	−0.279	15	4.43	0.321	5	0.86	−1.071
6	15	33	12.31	0.901	20	8.70	1.697	6	1.7	−0.276
7	13	39	10.12	−0.865	20	6.51	0.043	10	2.03	−1.478

n, number of sequences; S_{nuc}, number of variable nucleotide sites; S_{aa}, number of variable amino acid sites; S_{syn}, number of variable synonymous nucleotide sites; K_{nuc}, mean number of pairwise differences between nucleotide sequences; K_{aa}, mean number of pairwise differences between amino acid sequences; K_{syn}, mean number of pairwise differences at synonymous nucleotide sites; D_{nuc}, value of the test statistic for goodness-of-fit to a neutral model (20) (nucleotide sequences); D_{aa}, value of the test statistic for goodness-of-fit to a neutral model (amino acid sequences); D_{syn}, value of the test statistic for goodness-of-fit to a neutral model (20) (synonymous nucleotide sites).

systems, Mountain View, CA) using the GNU C compiler with an input tree obtained by the unweighted pair group mean method. A constant population size was assumed, reflecting the lack of substantial change in viral load during the study period.

RESULTS

The numbers of variable sites and mean pairwise divergence for each of the five samples obtained from the V3 region, along with the value obtained for the test statistic D (20), are presented (Table 1). The data were analyzed separately as nucleotide, amino acid, and synonymous nucleotide sites (estimated by an unweighted pathway method). For no sample was the pattern of sequence variation significantly different from that expected under neutrality. Analyzed as amino acid sequences, two of the five samples (years 4 and 6) showed a higher value of D that approached significance, suggesting that selection may be acting on the amino acid sequence without significantly affecting the overall pattern of variation at the nucleotide level.

In view of the prior suggestion that the V3 loop itself is the target of selection (3, 10, 25), the test was repeated using the same samples but considering separately the 35-codon sequence of the loop and the 42-codon sequence flanking it. The number of variable sites and mean number of pairwise differences are given in Table 2, again for both nucleotide and amino acid sequences as well as synonymous sites. For the V3 loop, one sample (year 6) showed a significant departure from the expectations of the neutral model on nucleotide sequence data ($P < 0.05$), and two samples (years 4 and 6) departed highly significantly when amino acid sequences were considered ($P < 0.001$), all due to relatively high pairwise divergence (K) for the number of variable sites. In contrast, no samples showed any departure from neutrality in the flanking regions either as nucleotide or as amino acid sequences. For no sample did the value of D for synonymous nucleotide variation depart significantly from that expected under neutrality.

The tests on V3 region sequences did not support the proposal that mutation–selection balance (which would be detected as a negative value of the test statistic D) determines the level of variation. To determine the generality of this result, the same test was applied to a data set obtained by Cheynier *et al.* (11) of sequences of the V1/V2 region of *env* amplified from individual white pulps from the same spleen. This study found substantial differences between white pulps in the proviral populations present. Because of this heterogeneity, the individual subpopulations were tested separately for fit to a neutral model. One (LD6) out of the four tests showed a highly significant departure from neutrality (Table 3), which was in the direction predicted for mutation–selection balance, with relatively low diversity (K_{aa}) for the number of variable sites (S_{aa}). The other three tests did not approach significance.

These results, taken together, provide evidence that selection can be detected in HIV-1 *env* gene sequences, but its effects are not detectable at the nucleotide level except when the V3 loop is considered separately. I have therefore estimated $2N_e\mu$ (θ) using the maximum likelihood-based method of Kuhner *et al.* (24) from the V3 sequences obtained in each of the five plasma samples from patient p82 (Table 4). A recent analysis of HIV-1 mutation rates found that the *in vivo* point mutation rate was 2.4×10^{-5} (26). From these two figures, estimates of N_e can be obtained that range from 1.0 to 2.1×10^3 (Table 4). Also in Table 4, the mean nucleotide diversity, expected to be equal to $2N_e\mu$, is presented for each sample (3.2–5.6%). From these, N_e would be estimated to range from 5.1×10^2 to 1.1×10^3.

DISCUSSION

The analysis of nucleotide sequences from the V3 region of the *env* gene revealed no significant departures from neutrality. The test applied was designed to detect the effect of mutation–

Table 2. Genetic variability and departures from neutrality in two components of the V3 region

Sample, year	n	S_{nuc}	K_{nuc}	D_{nuc}	S_{aa}	K_{aa}	D_{aa}	S_{syn}	K_{syn}	D_{syn}	d, %
V3 loop alone, 105 bp											
3	15	10	3.07	−0.007	6	2.17	0.616	1	0.13	−1.182	2.8
4	11	13	5.2	0.751	7	3.73	2.275†	3	0.55	−1.585	5.0
5	23	13	3.46	−0.062	8	2.52	0.528	2	0.25	−1.202	3.3
6	15	18	8.25	1.991*	11	5.47	2.378†	3	1.07	0.475	7.9
7	13	21	5.68	−0.690	11	3.21	−0.382	5	1.26	−0.777	5.2
V3 flanking regions, 126 bp											
3	15	15	4.40	−0.184	13	4.30	0.297	1	0.53	1.481	3.5
4	11	8	2.69	−0.063	6	2.29	0.467	2	0.76	0.346	2.1
5	23	12	2.83	−0.044	7	1.91	0.022	4	1.06	−0.746	2.2
6	15	15	4.07	−0.470	9	3.23	0.626	4	1.14	−0.235	3.2
7	13	18	4.69	−0.812	9	3.31	0.553	7	1.95	−0.514	3.7

Symbols as for Table 1 except: d, mean of the pairwise nucleotide distances for each sample.
*$P < 0.05$; †$P < 0.001$.

Table 3. Test for neutrality on amino acid sequence data of the V1/V2 region of *env* from individual white pulps from the same spleen (90 amino acids)

Sample	n	S_{aa}	K_{aa}	D_{aa}	d_{aa}, %
L-D6	20	21	2.64	−2.114*	3.4
L-D7	21	28	6.61	−0.582	8.1
L-D8	20	14	3.85	−0.089	4.8
L-D10	20	19	6.36	0.709	8.0

Symbols as for Table 1, except: d_{aa}, mean pairwise number of differences in amino acid sequence.
*$P < 0.05$.

selection balance, which has been invoked to explain under a deterministic model the low level of synonymous nucleotide variation observed in HIV-1 sequence data (13). The failure to detect any evidence of mutation–selection balance in nucleotide sequences raises questions about the appropriateness of such models.

When the sequences of the 35-amino acid V3 loop and its flanking regions were considered separately, the flanking sequences again showed no departure from neutrality when considered at the nucleotide or amino acid level. Sequences of the V3 loop itself departed significantly in two samples (years 4 and 6; Table 2), revealing a greater diversity than expected for the number of variable sites. The sequences present in those samples (21) show variation that includes a charge difference at amino acid site 11 in the V3 loop, known to be associated with differences in cell tropism (27–29). This supports the view that selection for differences in cell tropism can have a dominant effect on the viral population in some samples. In an earlier analysis of these data, the turnover of amino acid variants was described in these samples using a phylogenetic approach. The two sets of results suggest that strong selective forces act at the amino acid level in these sequences but that these forces are transient or stochastic in nature. Such "random environment" effects (30) would be expected to expose linked nucleotide variation to strongly stochastic influences, which could have a substantial impact on nucleotide sequence variation in this gene.

The method used here to evaluate departures from selective neutrality is based on the infinite sites model. It has been reported that the confidence intervals for the test statistic may be affected by variation in substitution rates over sites (31), resulting in a shift to more positive values of D even under neutrality. This would lead to type 1 errors with respect to positive values and type 2 errors with respect to negative values of D and imply that even the two significant departures in the V3 loop data set (Table 2) might not be real whereas the significance of the large negative value observed in the V1/V2 data (Table 3) would be increased. With the results obtained, there remains no evidence that mutation–selection balance has a significant effect on nucleotide variation in these sequences. Given that, it is desirable to identify the magnitude of the stochastic influences on the nucleotide sequences, and this is conventionally expressed as the effective population number (N_e). The effective population number estimates from these data are several orders of magnitude lower than the census size (Table 4). Although variation in substitution rates over sites also can cause errors in the estimation of θ, these are unlikely to be greater than 2- or 3-fold (24, 32).

Table 4. Estimates of θ, N_e, and nucleotide diversity d from plasma viral V3 sequences from patient p82

Sample, year	n	θ	N_e	d, %
3	15	0.0964	1.9×10^3	3.2
4	11	0.0576	1.2×10^3	3.5
5	23	0.0505	1.0×10^3	2.8
6	15	0.1052	2.1×10^3	5.6
7	13	0.0915	1.8×10^3	4.6

θ was estimated using the method of Kuhner *et al.* (24). d, mean of the pairwise nucleotide distances for each sample.

The census size for the viral population in this patient was estimated to be $\approx 10^8$, and estimates for other patients have ranged even higher (14, 15). The difference between N_e and N is so large that several factors probably contribute toward it. The titers of infectious viruses have been shown to be much lower than the titers of viral particles (33, 34), and a significant proportion of newly replicated viral genomes are expected to carry an inactivating mutation that will prevent their replication. In addition, it is possible that clonal amplification of infected T cell populations (11) could increase the variance of progeny number and further reduce N_e. Selection at other sites in the genome ("background" natural selection) has been shown to have a significant effect on N_e (35). However, the scale of the effect, with parameter values appropriate for HIV, is likely to be less than an order of magnitude. In addition, HIV infections are initiated from a small inoculum (36, 37) and increase very rapidly to $\approx 10^{10}$ in the first stages of infection (38), so a considerable reduction in N_e would be expected to be due to this expansion. All of these factors could act together with fluctuating selective forces to generate stochasticity in the evolution of viral sequences.

The low estimate of the effective population number of HIV implies that chance effects will have a significant impact on the evolution of the population, especially for lower values of s. We may conclude that the evolution of HIV is best described by finite population models (39). This conclusion may explain, for example, the persistence of zidovadine resistance-associated mutations within a population after therapy has been discontinued (40) or on transmission to a drug-naive patient (41, 42) although s for the resistant variants has been estimated to be over 10% in the absence of drug.† This result is very different from the *in vivo* estimate of 1% obtained by Goudsmit *et al.* (43) based on a deterministic model.

How much bigger is the global effective population size than the within-patient effective size? The ratio of nucleotide diversity estimates obtained within and between patients should reflect the ratio of the effective population size. Difficulties arise in making these comparisons because of the rapid saturation of synonymous substitutions, but estimates of nucleotide diversity for the V3 region between independent subtype B isolates lie between two and four times the value observed within many patients (44). This argument suggests that the subtype B effective population size would not be much greater than 10^4, reflecting the short time since the epidemic in the United States originated and probably a high variance in the number of transmissions from each infected patient. In a discussion based on analysis of the structure of their phylogenetic trees, Nee *et al.* (45) also have concluded that the global populations of both HIV and hepatitis C virus are a long way from evolutionary equilibrium. The effective population size for other viruses may therefore also be low.

†Harrigan, R., Bloor, S. & Larder, B., Fifth International Workshop on HIV Drug Resistance, July 3–5, 1996, Whistler, British Columbia, Canada.

I thank J. Felsenstein and M. Kuhner for software and the referees for constructive criticism of the manuscript. This work was supported by the Medical Research Council.

1. Wain-Hobson, S. (1993) *Curr. Opin. Genet. Dev.* **3,** 878–883.
2. Leigh Brown, A. J. & Holmes, E. C. (1994) *Annu. Rev. Ecol. Syst.* **25,** 127–165.
3. Seillier-Moiseiwitsch, F., Margolin, B. H. & Swanstrom, R. (1994) *Annu. Rev. Genet.* **28,** 559–596.

4. Goudsmit, J., Debouck, C., Meloen, R. H., Smit, L., Bakker, M., Asher, D. M., Wolff, A. V., Gibbs, C. J., Jr. & Gajdusek, D. C. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 4478–4482.
5. Rusche, J. R., Javaherian, K., McDanal, C., Petro, J., Lynn, D. L., Grimaila, R., Langlois, A., Gallo, R. C., Arthur, L. O., Fischinger, P. J., Bolognesi, D. P., Putney, S. D. & Matthews, T. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 3198–3202.
6. Takahashi, H., Merli, S., Putney, S. D., Houghten, R., Moss, B., Germain, R. N. & Berzofsky, J. A. (1989) *Science* **246,** 118–121.
7. Chesebro, B., Nishio, J., Perryman, S., Cann, A., O'Brien, W., Chen, I. S. & Wehrly, K. (1991) *J. Virol.* **65,** 5782–5789.
8. Hwang, S. S., Boyle, T. J., Lyerly, H. K. & Cullen, B. R. (1991) *Science* **253,** 71–74.
9. Simmonds, P., Zhang, L. Q., McOmish, F., Balfe, P., Ludlam, C. A. & Leigh Brown, A. J. (1991) *J. Virol.* **65,** 6266–6276.
10. Bonhoeffer, S., Holmes, E. C. & Nowak, M. (1995) *Nature (London)* **376,** 125.
11. Cheynier, R., Henrichwark, S., Hadida, F., Pelletier, E., Oksenhendler, E., Autran, B. & Wain-Hobson, S. (1994) *Cell* **78,** 373–387.
12. Nowak, M. A., May, R. M. & Anderson, R. M. (1990) *AIDS* **4,** 1095–1103.
13. Coffin, J. M. (1995) *Science* **267,** 483–489.
14. Ho, D. D., Moudgil, T. & Alam, M. (1989) *N. Engl. J. Med.* **321,** 1621–1625.
15. Piatak, M. Jr., Saag, M. S., Yang, L. C., Clark, S. J., Kappes, J. C., Luk, K. C., Hahn, B. H., Shaw, G. M. & Lifson, J. D. (1993) *Science* **259,** 1749–1754.
16. Kellam, P., Boucher, C. A., Tijnagel, J. M. & Larder, B. A. (1994) *J. Gen. Virol.* **75,** 341–351.
17. Boucher, C. A., O'Sullivan, E., Mulder, J. W., Ramautarsing, C., Kellam, P., Darby, G., Lange, J. M., Goudsmit, J. & Larder, B. A. (1992) *J. Infect. Dis.* **165,** 105–110.
18. Cleland, A., Watson, H. G., Robertson, P., Ludlam, C. A. & Leigh Brown, A. J. (1996) *J. Acquired Immune Defic. Syndr.* **12,** 6–18.
19. Wright, S. (1969) *Evolution and the Genetics of Populations Vol. 2 The Theory of Gene Frequencies* (Univ. of Chicago Press, Chicago), p. 512.
20. Tajima, F. (1989) *Genetics* **123,** 585–595.
21. Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A. & Leigh Brown, A. J. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4835–4839.
22. Wolfs, T. F. W., de Jong, J. J., Van den Berg, H., Tijnagel, J. M., Krone, W. J. & Goudsmit, J. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 9938–9942.
23. Wolfs, T. F. W., Zwart, G., Bakker, M. & Goudsmit, J. (1992) *Virology* **189,** 103–110.
24. Kuhner, M. K., Yamato, J. & Felsenstein, J. (1995) *Genetics* **140,** 1421–1430.
25. Leigh Brown, A. J. (1991) in *AIDS '91: A Year in Review*, eds. Adler, M. W., Gold, J. W. M. & Levy, J. A. (Current Science, London), pp. S35–S42.
26. Mansky, L. M. & Temin, H. M. (1995) *J. Virol.* **69,** 5087–5094.
27. Fouchier, R. A., Groenink, M., Kootstra, N. A., Tersmette, M., Huisman, H. G., Miedema, F. & Schuitemaker, H. (1992) *J. Virol.* **66,** 3183–3187.
28. Chesebro, B., Wehrly, K., Nishio, J. & Perryman, S. (1992) *J. Virol.* **66,** 6547–6554.
29. Milich, L., Margolin, B. & Swanstrom, R. (1993) *J. Virol.* **67,** 5623–5634.
30. Gillespie, J. H. (1992) *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford).
31. Bertorelle, G. & Slatkin, M. (1995) *Mol. Biol. Evol.* **12,** 887–892.
32. Tajima, F. (1996) *Genetics* **143,** 1457–1465.
33. Ho, D. D., Moudgil, T. & Alam, M. (1989) *N. Engl. J. Med.* **321,** 1621–1625.
34. Coombs, R. W., Collier, A. C., Allain, J.-P., Nikora, B., Leuther, M., Gjerset, G. & Corey, L. (1989) *N. Engl. J. Med.* **321,** 1626–1631.
35. Nordborg, M., Charlesworth, B. & Charlesworth, D. (1996) *Genet. Res.* **67,** 159–174.
36. Zhang, L. Q., MacKenzie, P., Cleland, A., Holmes, E. C., Leigh Brown, A. J. & Simmonds, P. (1993) *J. Virol.* **67,** 3345–3356.
37. Zhu, T., Mo, H., Wang, N., Nam, D. S., Cao, Y., Koup, R. A. & Ho, D. D. (1993) *Science* **261,** 1179–1181.
38. Katzenstein, T. L., Pedersen, C., Nielsen, C., Lundgren, J. D., Jakobsen, P. H. & Gerstoft, J. (1996) *AIDS* **10,** 167–173.
39. Kelly, J. K. (1994) *Genet. Res.* **64,** 1–9.
40. Boucher, C. A., van Leeuwen, R., Kellam, P., Schipper, P., Tijnagel, J., Lange, J. M. A. & Larder, B. A. (1993) *Antimicrob. Agents Chemother.* **37,** 1525–1530.
41. Conlon, C. P., Klenerman, P., Edwards, A., Larder, B. A. & Phillips, R. E. (1994) *J. Infect. Dis.* **169,** 411–415.
42. Erice, A., Mayers, D. L., Strike, D. G., Sannerud, K. J., McCutchan, F. E., Henry, K. & Balfour, H. (1993) *N. Engl. J. Med.* **328,** 1163–1165.
43. Goudsmit, J., de Ronde, A., Ho, D. D. & Perelson, A. S. (1996) *J. Virol.* **70,** 5662–5664.
44. Korber, B. T., MacInnes, K., Smith, R. F. & Myers, G. (1994) *J. Virol.* **68,** 6730–6744.
45. Nee, S., Holmes, E. C., May, R. M. & Harvey, P. H. (1994) *Philos. Trans. R. Soc. London B* **344,** 77–82.